A Probabilistic Model for Dimensionality Reduction in Information Retrieval and Filtering

Chris H.Q. Ding

NERSC Division, Lawrence Berkeley National Laboratory

University of California, Berkeley, CA 94720. chqding@lbl.gov

January 2001 (updated November 2001)

Abstract

Dimension reduction methods, such as Latent Semantic Indexing (LSI), when applied to semantic spaces built upon text collections, improve information retrieval, information filtering and word sense disambiguation. A new dual probability model based on similarity concepts is introduced to explain the observed success. Semantic associations can be quantitatively characterized by their statistical significance, the *likelihood*. Semantic dimensions containing redundant and noisy information can be separated out and should be ignored because their contribution to the overall statistical significance is negative, giving rise to LSI: LSI is the optimal solution of the model. The peak in likelihood curve indicates the existence of an intrinsic semantic dimension. The importance of LSI dimensions follows the Zipf-distribution, indicating that LSI dimensions represent latent concepts. Document frequency of words follow the Zipf distribution, and the number of distinct words follows log-normal distribution. Experiments on four standard document collections both confirm and illustrate the results and concepts presented here.

Keywords: Latent Semantic Indexing, intrinsic semantic subspace, dimension reduction, word-document duality, Zipf-distribution.

1 Introduction

As computers and the internet become a part of our daily lives, effective and automatic information retrieval and filtering methods become essential to deal with the explosive growth of accessible information. By far, most information is text based. Many current systems, such as Internet search engines, retrieve information by exactly matching query keywords to words indexing the documents in the database. A well-known problem[Furnas et al, 1987] is the ambiguity in word choices. One searches for "car" related items while missing items relating to "auto" (synonyms problem). One looks for "capital" city, and gets venture "capital" instead (polysemy problem). Although "land preserve" and "open space" express very similar ideas, searching on Web search engines will retrieve two very different sets of webpages with little overlap. These kinds of problems are well-known.

One solution is to manually classify information into different categories using human judgements. This categorized or filtered information, in essence, reduce the size of the relevant information space and is thus more convenient and useful.

A somewhat similar, but automatic approach is to use dimension reduction (data reduction) methods such as the Latent Semantic Indexing (LSI) method [Deerwester et al, 1990, Dumais, 1995, Berry et al 1995]. LSI automatically computes a much smaller semantic subspace from the original text collection, which improves recall and precision in information retrieval [Deerwester et al, 1990, Bartell et al,1995, Zha et al, 1998, Hofmann, 1999, Husbands et al, 2000], information filtering or text classification[Dumais, 1995, Yang, 1995, Baker and McCallum, 1998], and word sense disambiguation [Schütze, 1998].

The effectiveness of LSI in these empirical studies is often attributed to reduction of noise, redundancy, and ambiguity. Synonyms and polysemy problems are somehow reduced in the process. Several recent studies [Papadimitriou et al, 1998, Hofmann, 1999, Dhillon and Modha,2001, Zha et al, 1998, Bartell et al,1995, Story, 1996] shed some lights on LSI method (see section 9 for detailed discussions).

A central question, however, remains unresolved. Since LSI is a pure "numerical" and automatic procedure, the noisy and redundant semantic information must be associated with a numerical quantity that was reduced or minimized in LSI. But how can one define a quantitative measure for semantic information? How can one verify that this quantitative

measure of semantic information is actually reduced or minimized in LSI?

In this paper we address this question by introducing a new probabilistic model based on document-document and word-word similarities, and show that LSI is the optimal solution of the model.

Furthermore, we use the statistical significance, i.e., the *likelihood*, as a quantitative measure of the LSI semantic dimensions. We calculated likelihood curves of four standard document collections; as LSI subspace dimension k increases (Figure 1), all likelihood curves arise very sharply at the beginning, gradually turn into a convex peak, and decrease steadily afterwards. This unambiguously demonstrated that the dimensions after the peak contain no statistically meaningful information — they represent noisy and redundant information. The existence of the semantic subspace that contains the maximum statistically meaningful information, the "intrinsic semantic subspace", explains the observed performance improvements for LSI.

Our model indicates that the statistical significance of a LSI dimension is related to the square of its singular value. On the four document collections, the statistical significance of LSI dimensions are found to follow a Zipf-law, indicating LSI dimensions represent latent concepts in the same way as webpages, cities, and English words do.

We further studied the document frequency distribution which helps to explain the Zipf-law characteristics of LSI dimensions. We also investigated the distribution of distinct words which reveal some internal structure of document collections. Overall, our studies provides a statistical framework for understanding and further developing of LSI/SVD type dimension reduction methods.

The rest of the paper is organized as follows. In section 2, we briefly outline the vector space model of information retrieval to provide necessary context and notations. In section 3, LSI is explained and some fundamental questions are raised. The starting point of the probablistic model is similarity matrices which are discussed in Section 4. The dual probability model is outlined in detail in section 5. Intrinsic semantic subspace are calculated for four collections in section 6. In section 7, the statistical significance of LSI dimensions is found to follow Zipf-law. Distributions of document frequency and distinct words are examined in section 8. Using the context and notations already established, related works are discussed

in some detail in section 9. Discussions on invariance properties, normalization factors, and separation of words and documents in LSI space are given in section 10. Conclusions are made in section 11. Preliminary results of this work has been presented in [Ding, 1999].

2 Semantic Vector Space

One of the fundamental relationships in human language is the dual relationship, the mutual interdependence, between words and concepts. Concepts are expressed by our choices of words, while the meanings of words are inferred by their usage in different contexts. Casting this relationship in mathematical form is the semantic vector space [Salton and McGill, 1983]. A document (title plus abstract or first few paragraphs) is represented by a vector in a linear space indexed by d words (word vector space). Similarly, a word (which has clear content and is often stemmed, also commonly referred to as *term*) is represented by a vector in a linear space spanned by n document/contexts (document vector space). These dual representations are best captured by the word-to-document association matrix X, where each column $\mathbf{x}_i$ represents a document[1], and each row $\mathbf{t}^\alpha$ represents a word (term)[1]:

$$X = \begin{pmatrix} x_1^1 & \dots & x_n^1 \\ \vdots & \ddots & \vdots \\ x_1^d & \dots & x_n^d \end{pmatrix} \equiv (\mathbf{x}_1 \cdots \mathbf{x}_n) \equiv \begin{pmatrix} \mathbf{t}^1 \\ \vdots \\ \mathbf{t}^d \end{pmatrix} \tag{1}$$

where $x_i^\alpha \equiv (\mathbf{x}_i)^\alpha \equiv (\mathbf{t}^\alpha)_i$.

The matrix element x_i^α contains the term frequence (`tf`) of term α occurring in the document i, properly weighted by other factors[Salton and Buckley, 1988], most often the inverse document frequency (idf),

$$x_i^\alpha = \mathtt{tf}_i^\alpha \cdot \log(n/\mathtt{df}^\alpha) \tag{2}$$

where the document frequence $\mathtt{df}^\alpha$ is the number of documents the word α occurs in. (Following the IR convention, we use `tf` to denote term frequency and `df` to denote document

[1]In this paper, capital letters refer to matrices, bold face lower-case letters to vectors: vectors with subscript represent documents and vectors with superscript represent terms; α, β sum over all d terms and i, j sum over all n documents.

frequency, and use them directly in mathematical expressions.) Note that most matrix elements of X are zero because an index word usually occurs in only a few documents (see section 8).

Information retrieval on the document collection is typically handled by keywords matching. A user query $\mathbf{q}$, consisting of a set of keywords (terms), is treated as a document. The keyword matching is equivalent to a dot-product between the query vector and a document vector (variable document length is accounted for by normalizing document vectors to unit length). *Relevance* scores for each of the n documents form a row vector, and are calculated as

$$\mathbf{s} = \mathbf{q}^T X.$$

Documents are then sorted according to their relevance score and returned to user.

Information filtering (also called as text categorization), such as classifying an incoming news item or email into predefined categories, can be done in a number of ways. A simple method is to calculate a centroid vector $\mathbf{c}_k$ of category k, i.e., the average of all documents in the category[Dumais, 1995]. All m centroid vectors for m categories form a $d\times m$ matrix $C = (\mathbf{c}_1 \cdots \mathbf{c}_m)$. Another method is to solve for mapping vectors $\mathbf{c}_k$ so that[Yang, 1995]

$$C = \arg\min_C ||C^T X - B||.$$

In the least square problem, the $m\times n$ matrix B defines categories for each document. The Frobenius norm of a matrix J is defined as $||J||^2 = \sum_{i=1}^{n}\sum_{k=1}^{m}(J_i^k)^2$. A new incoming document is then projected onto these centroids or mapping vectors to get similarity scores for each category with appropriate thresholding. Note that a document may belong to several categories.

In word sense disambiguation[Schütze, 1998], the calculation involves a *collocation* matrix, very similar to the term-document matrix X. Here, each column represent a context for a target word, say *capital*. The column contains the words that collocate with the target word within a text window of a fixed number of words or a sentence. The index words of the context space are the same as in the term-document matrix. These contexts are then clustered to find different senses of the target word (the word "capital" could mean "capital goods," "city where state government seats," etc).

3 Dimension Reduction: Latent Semantic Indexing

In the *initial* semantic vector space, the word-document relations contain redundancy, ambiguity, and noise — the subspace containing meaningful semantic associations is much smaller than the initial space. One method to achieve this is to perform a dimension reduction (data reduction) to a semantic subspace that contains essential and meaningful associative relations.

LSI is one such dimension reduction method. It automatically computes a *subspace* containing meaningful semantic associations which is much smaller than the initial space. This is done through the singular value decomposition (SVD) [Golub and Van Loan, 1989] of the term-document matrix:

$$X = \sum_{k=1}^{r} \mathbf{u}_k \sigma_k \mathbf{v}^k = (\mathbf{u}_1 \cdots \mathbf{u}_r) \begin{pmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_r \end{pmatrix} \begin{pmatrix} \mathbf{v}^1 \\ \vdots \\ \mathbf{v}^r \end{pmatrix}, \tag{3}$$

where r is the rank of the matrix X, $\Sigma_r \equiv \mathrm{diag}(\sigma_1 \cdots \sigma_r)$ are the singular values, $U_r \equiv (\mathbf{u}_1 \cdots \mathbf{u}_r)$ and $V_r \equiv (\mathbf{v}^1 \cdots \mathbf{v}^r)$ are left and right singular vectors. Typically the rank r is order of $min(d, n)$, which is about 10,000. However, if we truncate the SVD, i.e., keep only the first k largest terms, the resulting

$$X = U_r \Sigma_r V_r^T \simeq U_k \Sigma_k V_k^T$$

is a good approximation. Note here, $k \sim 200$ and $r \sim 10,000$, thus a very substantial dimensionality reduction. (Good illustrative examples were given in [Deerwester et al, 1990, Berry et al 1995].)

In a LSI k-dim subspace, a document $\mathbf{x}_i$ is represented as its projection in the subspace, $U_k^T \mathbf{x}_i$, and all n documents $(\mathbf{x}_1 \cdots \mathbf{x}_n)$ are represented as $U_k^T X = \Sigma_k V_k^T$. Queries and mapping vectors are transformed in the same way as documents. Therefore, in text retrieval, the score vector in LSI subspace is evaluated as $\mathbf{s} = (U_k^T \mathbf{q})^T (U_k^T X) = (\mathbf{q}^T U_k)(\Sigma_k V_k^T)$. Using these LSI dimensions in document retrieval, both recall and precision are improved [Deerwester et al, 1990, Bartell et al,1995, Zha et al, 1998, Hofmann, 1999, Husbands et al, 2000, Caron, 2000].

The greatly reduced dimension also reduces the complexity and noise in text categorizations. For example, the centroid matrix or mapping matrix is reduced from $d \times m$ to $k \times m$,

a very substantial reduction of degree of freedom [Dumais, 1995, Yang, 1995]. The same dimension reduction is also effectively used in Naive Bayes categorization[Baker and McCallum, 1998]. All these lead to more accurate categorization results.

In word sense discrimination, the LSI dimension reduction is necessary to reduce the computational complexity in the clustering process to find different word senses, thus leading to better clustering results and better disambiguation results [Schütze, 1998].

The success of LSI is attributed to that the LSI subspace captures the essential associative semantic relationships better than the original document space, and thus partially resolves the word choice (synonyms) problem in information retrieval, and redundant semantic relationships in text categorization.

Mathematically, LSI with a truncated SVD is the best approximation of X in the reduced k-dim subspace (Eckart-Young Theorem, see [Golub and Reinsch, 1971]). However, the *improved* results in information retrieval and filtering indicates the LSI goes beyond mathematical approximations.

From statistical point of view, LSI amounts to an effective dimensionality reduction, similar to principal component analysis in statistics. Dimensions with small singular values are often viewed as representing semantic noises and thus are ignored. This generic argument, considering its fundamental importance, needs to be clarified. For example, how small do the singular values have to be in order for the dimensions to be considered noise? A small singular value only indicates that the corresponding dimension is not as important as those with large singular values, but "less important" in itself does not directly imply "noise" or "redundancy".

Thus the question becomes how to quantitatively characterize and measure the associative semantic relationships. If we have a quantitative measure, we can proceed to verify if dimensions with smaller singular values do indeed represent noise.

Directly assigning an appropriate numerical score for each associative relationship appears to be intangible. Instead, we approach the problem with a probabilistic model, and use statistical significance, the *likelihood*, as the quantitative measure for the semantic dimensions. The governing relationship in the probabilistic model is similarity between documents and between words, which we discuss next.

4 Similarity Matrices

It is generally accepted that the dot-product between two document vectors (normalized to 1 to account for different document lengths) is a good measure of the correlation or similarity of word usages in the two documents; therefore the similarity between the two documents is defined as

$$\mathrm{sim}(\mathbf{x}_1, \mathbf{x}_2) = \mathbf{x}_1 \cdot \mathbf{x}_2.$$

X^TX contains similarities between all pairs of documents, and is the similarity matrix between documents.

Similarly, the dot-product between two word vectors

$$\mathrm{sim}(\mathbf{t}^1, \mathbf{t}^2) = \mathbf{t}^1 \cdot \mathbf{t}^2$$

measures their co-occurrences through all documents in the collection, and therefore their closeness or similarity. XX^T contains similarities between all pairs of words and is the word-word similarity matrix. If we assign binary weights to term-document matrix elements x_i^α, one can easily see that XX^T contains the word-word co-occurrence frequency when the context window size is set to the document length. For other weighting, such as the tf.idf weighting, XX^T is also a good measure of the word co-occurrence.

These similarity matrices define the semantic relationships, and are of fundamental importance in information retrieval [Salton and McGill, 1983]. Note that document-document similarity is defined in the word-vector-space, while word-word similarity is defined in document-vector-space. This strong dual relationship between documents and words is a key feature of our model.

5 Dual Probability Model

Traditional IR probabilistic models [Rijsbergen, 1979, Fuhr, 1992] focus on relevance to queries. Our approach focuses on the data, the term-document association matrix X. Query-specific information is ignored at present, but may be included in future developments.

Documents are data entries in the d-dimensional term-space (index space), and they do not occur randomly. We assume they are generated according to certain probability

distributions. To find the generative probability, we assume (1) The probability distribution is governed by k characteristic (normalized) document vectors $\mathbf{c}_1 \cdots \mathbf{c}_k$ (collectively denoted as C_k), which will be later identified as LSI dimensions; (2) The occurrence of a document $\mathbf{x}_i$ is proportional to its similarity to $\mathbf{c}_1 \cdots \mathbf{c}_k$. When projecting onto a dimension $\mathbf{c}_j$, $\pm$ signs are equivalent, thus we use $(\mathbf{c}_j \cdot \mathbf{x})^2$ instead of $\mathbf{c}_j \cdot \mathbf{x}$; (3) $\mathbf{c}_1 \cdots \mathbf{c}_k$ are statistically independent factors; (4) Their contribution to total probability for a document is additive. With these assumptions, and further motivated by Gaussian distribution, we consider the following generative model:

$$\Pr(\mathbf{x}_i|\mathbf{c}_1 \cdots \mathbf{c}_k) = e^{(\mathbf{x}_i \cdot \mathbf{c}_1)^2} \cdots e^{(\mathbf{x}_i \cdot \mathbf{c}_k)^2} / Z(C_k) \tag{4}$$

where $Z(C_k)$ is the normalization constant (also called partition function). The maximum likelihood estimation(MLE) method is used to obtain $\mathbf{c}_1 \cdots \mathbf{c}_k$ as the optimal parameters for the probability model. Note that assumption (3) requires $\mathbf{c}_1 \cdots \mathbf{c}_k$ to be mutually orthogonal. Assuming $\mathbf{x}_i$ are independently, identically distributed, the log-likelihood of the model is calculated as

$$\ell(C_k) \equiv \log \prod_{i=1}^{n} \Pr(\mathbf{x}_i|\mathbf{c}_1 \cdots \mathbf{c}_k)$$

which is

$$\ell(C_k) = \sum_{i=1}^{n} [\sum_{j=1}^{k} (\mathbf{x}_i \cdot \mathbf{c}_j)^2 - \log Z(C_k)] = \sum_{j=1}^{k} \mathbf{c}_j^T X X^T \mathbf{c}_j - n \log Z(C_k) \tag{5}$$

after some algebra and using the relation

$$\sum_{i=1}^{n} (\mathbf{x}_i \cdot \mathbf{c})^2 = \sum_{\alpha,\beta=1}^{d} c^\alpha (X X^T)^{\alpha\beta} c^\beta \tag{6}$$

for any given $\mathbf{c} = \mathbf{c}_j$. Note that in Eq.(5), it is the word-word similarity matrix XX^T (the word co-occurrence matrix) that arises here as a natural consequence of MLE, rather than the document-document similarity matrix that one might have expected. This is because of the dual relationship between documents and words. Rephrasing it differently, documents are data points which live in the index space (word space). XX^T measures the "correlation" between components of data points, i.e., correlation between words. When properly normalized, XX^T would not change much if more data points are included, thus serving a role similar to the covariance matrix in principal component analysis. Therefore, understanding document relationship is ultimately related to the understanding of word co-occurrence. This is an interesting result of our model.

In general, finding C_k that maximizes $\ell(C_k)$ involves a rather complicated numerical procedure, particularly due to the analytically intractable $Z(C_k)$ as a high ($d = 10^3 - 10^5$) dimensional integral,

$$Z_k = \int \cdots \int e^{(\mathbf{x}\cdot\mathbf{c}_1)^2 + \cdots + (\mathbf{x}\cdot\mathbf{c}_k)^2} dx^1 \cdots dx^d.$$

However, note that $n\log Z(C_k)$ is a very slow changing function in comparison to $\sum_j \mathbf{c}_j^T XX^T \mathbf{c}_j$: (1) In essence, $\mathbf{c}_j$ is similar to the mean vector μ in Gaussian distribution, where the normalization constant is independent of μ. Thus $Z(C_k)$ should be nearly independent of $\mathbf{c}_j$. (2) The logarithm of a slow changing function changes even slower. Thus $n\log Z(C_k)$ can be regarded as fixed, and we concentrate on maximizing the first term in Eq.(5).

The symmetric positive-definite matrix XX^T has a spectral decomposition (eigenvector expansion) with non-negative eigenvalues only: $XX^T = \sum_{i=1}^r \lambda_i \mathbf{u}_i \mathbf{u}_i^T$, $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_r \geq 0$, here λ_i and $\mathbf{u}_i$ are the ith eigenvalue and eigenvector ($XX^T\mathbf{u}_i = \lambda_i \mathbf{u}_i$). Therefore the optimal solution for characteristic dimensions $\mathbf{c}_1 \cdots \mathbf{c}_k$ in maximizing $\sum_j \mathbf{c}_j^T XX^T \mathbf{c}_j$ are $\mathbf{u}_1 \cdots \mathbf{u}_k$. They are precisely the left singular vectors $\mathbf{u}_1 \cdots \mathbf{u}_k$ in SVD of X used in LSI. Thus LSI is the optimal solution of our model, and we will refer to $\mathbf{u}_1 \cdots \mathbf{u}_k$ as LSI dimensions. The final maximal likelihood is

$$\ell(U_k) = \lambda_1 + \cdots + \lambda_k - n\log Z(U_k). \tag{7}$$

The above analysis can be carried out for words in document vector space: we model words as defined by their occurrences in all documents. Here data points are words, indexed by documents and are represented as row vectors in the word-document matrix X. Consider k (normalized) row vectors $\mathbf{r}^1 \cdots \mathbf{r}^k$ (collectively denoted as R_k) representing k characteristic words. Using the word-word similarity, we assume the probability for the occurrence of word $\mathbf{t}^\alpha (\alpha = 1, \cdots, d)$ to be

$$\Pr(\mathbf{t}^\alpha | \mathbf{r}^1 \cdots \mathbf{r}^k) = e^{(\mathbf{t}^\alpha \cdot \mathbf{r}^1)^2} \cdots e^{(\mathbf{t}^\alpha \cdot \mathbf{r}^k)^2} / Z(R_k). \tag{8}$$

The log-likelihood becomes

$$\ell(R_k) \equiv \log \prod_{\alpha=1}^d \Pr(\mathbf{t}^\alpha | \mathbf{r}^1 \cdots \mathbf{r}^k) = \sum_{j=1}^k \mathbf{r}^{jT} X^T X \mathbf{r}^j - d\log Z(R_k), \tag{9}$$

after some algebra, and noting

$$\sum_{\alpha=1}^d t_i^\alpha t_j^\alpha = (X^T X)_{ij}. \tag{10}$$

The document-document similarity matrix X^TX arise naturally here. Here again we maximize the log-likelihood Eq.(8). Following the same line of reasoning for $\ell(C_k)$ of Eq.(5), we see that the second term, $d\log Z(R_k)$, is a slow changing function; thus we only need to maximize the first term in Eq.(9). The symmetric positive definite matrix X^TX has a spectral decomposition: $X^TX = \sum_{\alpha=1}^{r} \xi_\alpha (\mathbf{v}^\alpha)^T \mathbf{v}^\alpha$, $\xi_1 \geq \xi_2 \geq \cdots \geq \xi_r \geq 0$, here ξ_α and $\mathbf{v}^\alpha$ are the αth eigenvalue and eigenvector, $X^TX\mathbf{v}^\alpha = \xi_\alpha \mathbf{v}^\alpha$. The optimal solution for characteristic words $\mathbf{r}^1 \cdots \mathbf{r}^k$ in maximizing $\sum_j \mathbf{r}^{jT} X^TX \mathbf{r}^j$ are $\mathbf{v}^1 \cdots \mathbf{v}^k$. By construction of SVD, $\mathbf{v}^1 \cdots \mathbf{v}^k$ are precisely the right singular vectors of SVD of X. Therefore, $\mathbf{v}^1 \cdots \mathbf{v}^k$ of LSI are the optimal solution of the document-space model, and the maximal log-likelihood is

$$\ell(V_k) = \xi_1 + \cdots + \xi_k - n\log Z(V_k). \tag{11}$$

Eqs.(4,8) are dual probability representations of the LSI. This dual relation is further enhanced by the facts: (a) XX^T and X^TX have the same eigenvalues

$$\lambda_j = \xi_j = \sigma_j^2, \ j = 1, \cdots, k; \tag{12}$$

(b) left and right LSI vectors are related by

$$\mathbf{u}_j = (1/\sigma_j) X (\mathbf{v}^j)^T, \mathbf{v}_j = (1/\sigma_j)\mathbf{u}_j^T X. \tag{13}$$

Thus both probability representations have the same maximum log-likelihood

$$\ell_k = \sigma_1^2 + \cdots + \sigma_k^2 \tag{14}$$

up to a small and slowly changing normalization constant. This is the direct consequence of the dual relationship between words and documents. In particular, for statistical modeling of the observed word-text co-occurrence data, both probability models should be considered with the same number k, as is the case in the SVD.

Eq.(14) also suggests that the contribution (or the statistical significance) of each LSI dimension is approximately the square of its singular value. This quadratic dependence indicates that LSI dimensions with small singular values are much more insignificant than have been perceived earlier: previously it was generally thought that contributions of LSI dimensions are proportional to singular values linearly, since their singular values appear directly in SVD (cf. Eq.3). Suppose we have two LSI dimensions with singular values 10 and

1 respectively. Compared to the importance of the first dimension, the second dimension is only 1% (rather than 10%) as important. This result gives the first insight as to why one only needs to keep a small number of LSI dimensions and ignore the large number of dimensions with small singular values.

6 Intrinsic Semantic Subspace

The central theme in LSI is that the LSI subspace captures the essential meaningful semantic associations while reducing redundant and noisy semantic information. Our model provides a quantitative mechanism to verify this claim by studying the statistical significance of the semantic dimensions: If a few semantic dimensions can effectively characterize the data statistically, as indicated by the likelihood of the model, we believe they also effectively represent the semantic meanings/relationships as defined by the cosine similarity. We further conjecture that semantic dimensions with small eigenvalues contain statistically insignificant information, and their inclusion in the probability density will not increase the the likelihood. In LSI, they represent redundant and noisy semantic information.

Thus the key to resolving this central question relies on the behavior of the log-likelihood as a function of k. In the word-space model it is given by Eq.(7). The analytically intractable $Z(U_k) = Z(\mathbf{u}_1 \cdots \mathbf{u}_k)$ can be evaluated numerically by statistical sampling. We generate uniform random numbers in the domain of the integration: on the unit sphere in d-dimensional space, with all components positive. We found this sampling method converges very quickly. It achieves an accuracy of 4 decimal places with merely 2000 points for $d = 2000 - 5000$ dimensions.

The log-likelihood of LSI word dimensions (defined in document-space model) (cf. Eq.11) can be calculated similarly. Due to the fact that column vectors of X are normalized to 1, the matrix norm of X is $||X||^2 = n$. Thus the normalization of terms, i.e., row vectors, should be

$$||\mathbf{t}^\alpha||^2 = \sum_{i=1,n} (x_i^\alpha)^2 = n/d, \;\; \alpha = 1, \cdots, d, \tag{15}$$

and the domain of integration is the first quadrant on the sphere of radius $\sqrt{n/d}$ in n-dimensional document space.

Likelihood curves are calculated for four standard test document collections in IR: CRAN (1398 document abstracts on Aeronautics from Cranfield Institute of Technology), CACM (3204 abstracts of articles in *Communications of ACM*), MED (1033 abstracts from National Library of Medicine), and CISI (1460 abstracts from Institute of Scientific Information). In the term-document matrices we use the standard term frequency- inverse document frequency (tf.idf) weighting. The calculated likelihood curves are shown in Figure 1.

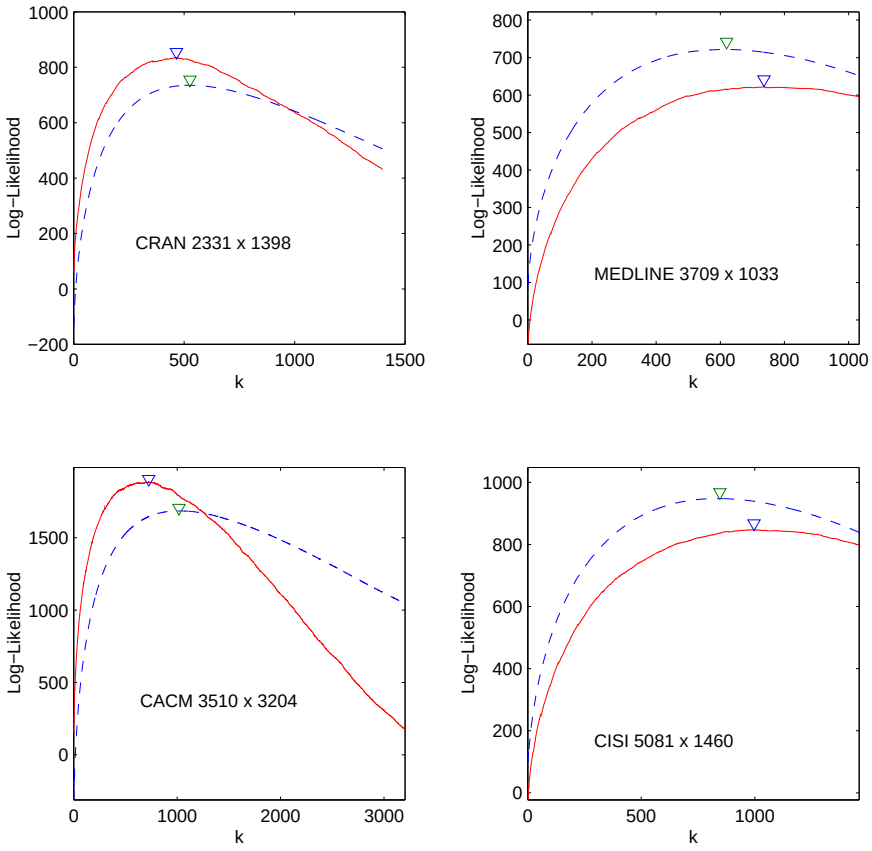

Figure 1: Log-likelihood curves for the four document collections. Solid lines for moduling documents in word space, $\ell(U_k)$, and dashed lines for moduling words in document space, $\ell(V_k)$. The intrinsic semantic subspace dimension, $k_{\text{int}}^{(u)}$ for word space and $k_{\text{int}}^{(v)}$ for document space, are also indicated. Number of words d and number of documents n for each collection is given after the collection name.

For all four collections, both word-space and document-space likelihoods grow rapidly and steadily as k increases from 1 up to k_{int} , clearly indicating that the probability models provide better and better statistical descriptions of the data. They reach a peak at k_{int} . However, starting from $k > k_{\text{int}}$, the likelihood decreases steadily, indicating no meaningful

statistical information is represented by those LSI dimensions with smaller eigenvalues. For all four collections, the intrinsic dimensions determined from document-space, $k_{\text{int}}^{(u)}$, and from word-space, $k_{\text{int}}^{(v)}$, are fairly close as indicated in Figure 1. All these indicate the correctness of the dual probability model.

Note that the theoretical k_{int} from the likelihood curve for CACM is quite close to that experimentally determined for text classification[Yang, 1995]. For Medline, however, $k_{\text{int}}^{(v)}$ is larger than the experimentally determined value [Deerwester et al, 1990, Zha et al, 1998], based on the 11-point average precision for 30 standard queries. Since the model contains no information on the queries, these reasonable agreements indicate that the statistical model and the statistical significance-based arguments capture some essential relationships involved.

Overall, the general trend for the four collection is quite clear. These likelihood curves quantitatively and unambiguously demonstrate the existence of an intrinsic semantic subspace: dimensions with small eigenvalues do represent redundant or noisy information, and contributes negatively to the statistical significance. This is one of the main results of this work.

7 Do LSI Dimensions Represent Concepts?

LSI dimensions are optimal solutions for the characteristic document vectors introduced in the dual probability model. Note that the similarity relationship, i.e., the dot-product, can also be viewed as projection of document $\mathbf{x}_i$ onto characteristic vector $\mathbf{c}_j$ (see Eq.4). Thus LSI dimensions are actually projection directions, as are obvious from the SVD point of view.

Besides the projection directions, do these LSI dimensions represent something about the document collection? Or equivalently, do the projection directions mean something? As explained in the original paper [Deerwester et al, 1990], the exact meaning of those LSI dimensions are complex and can not directly inferred (thus named "latent" semantic indexing).

Our probability model provides additional insights to this issue. The statistical significance or importance of each LSI dimension directly relates to its singular value squared (cf.

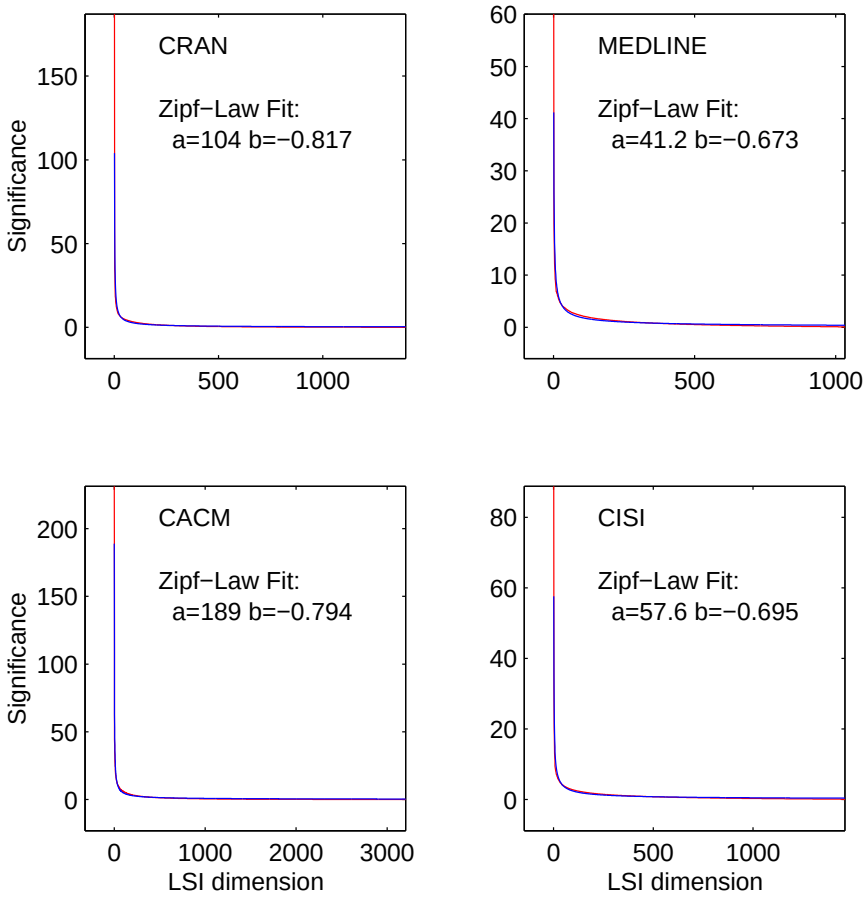

Figure 2: Statistical significances of the LSI/SVD dimensions for the four document collections. The Zipf-law with only two parameters fits the data extremely well: the original data and fit are essentially indistinguishable.

Eq.14). They are calculated for all four document collections and are shown in Figure 2.

The statistical significance of LSI dimensions clearly follow a Zipf law, $\sigma_i^2 = a \cdot i^b$, with the exponent b very close to -1. These fits are very good: the data and the fits are almost indistinguishable for all 4 collections. We conjecture that the Zipf-law is obeyed by singular values squared of most if not all document collections.

Zipf-law, named after the Harvard linguist George Zipf [Zipf, 1949], is the observation that frequency of occurrence f, as a function of the rank i, is a power-law function

$$f_i = a \cdot i^b \tag{16}$$

with the exponent b close to -1. There are a wide range of social phenomena that obey Zipf-law. The best known example is the frequency of word usage in English and other languages.

Ranking all cities in the world according to their population, they also follow Zipf-law. Most recently on the Internet, if we rank the website by their popularity, the number of user visits, the webpage popularity also obey the Zipf-law[Glassman, 1994].

One common theme among English words, cities, webpages, etc, is that each one has distinct characters or identities. Since LSI dimensions on all document collections display very clear Zipf-distribution, we may infer that LSI dimensions represent some latent concepts or identities in a similar manner as English words, cities, or webpages do.

8 Characteristics of Document Collections

To provide a perspective on the statistical approach discussed above, we further investigate the characteristics of the four document collections. There are many studies on statistical distributions in the context of natural language processing (see [Manning and H. Schütze, 1999] and references there). One of the emphasis there is the word frequency distribution in documents, ranging from the simple Poisson distribution to the K-mixture distribution [Katz, 1996]. Here we studied the two distributions that have close relations to the probablistic model discussed above.

8.1 Document Frequency Distribution

We first study the document frequency (`df`) of each word, i.e., the number of document a word occurs in. In the term-document matrix X, this corresponds to the number of nonzero elements in each row.

In Figure 3, we show the distribution of document frequency for the four collections. Plotted are the number of words at a given document frequency, the histogram. For all four collections, there are large number of words which have small `df`. For example, for CRAN, there are 466 words with `df`=2, 243 words with `df`=3, etc. On the other hand, the number of words with large `df`is small. There are only a total number of 186 words which has `df` $\geq$ 100, although `df` can reach as high as 860 for one word.

As one may expect from earlier discussions, this kind of distribution can be described by

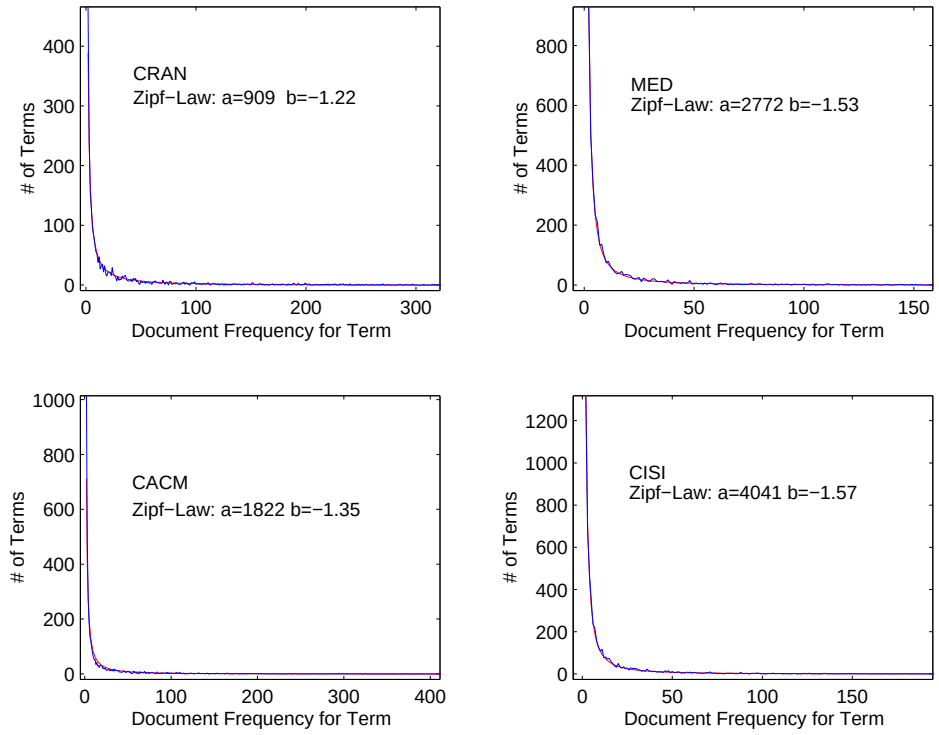

Figure 3: Distributions of document frequency for terms in the four collections. The Zipf-law fits are also shown.

the Zipf distribution (more precisely, the power law),

$$N(\mathtt{df}) = a \cdot \mathtt{df}^b$$

In fact, the Zipf-law fits data very well for all four collections, as shown in Figure 4. The exponents are generally close to -1. Zipf-law originally describe the frequency of word usage; Here we see that Zipf-law also governs the document frequence of *content* words.

The distribution of document frequency gives a better understanding of the LSI dimensions discussed in previous section. Since LSI dimensions are essentially linear combinations of the words, we may say that the Zipf-law behavior of the words directly implies that the statistical significance of the LSI dimensions also follow the Zipf-law. This analogy further strengthen our previous arguments that LSI dimensions represent latent concepts, in much the same way as the indexing words do.

In the literature, the average document frequency is often quoted as a numerical characterization of the document collection. For Gaussian type distributions, the mean (average) is the center of the bell-shaped curve and is a good characterization of the distribution;

however, for scale-free Zipf type of distributions, the mean does not capture the essential features of the distribution. Whether `df=1` words are included or not will change the mean quite significantly since they dominate the averaging process; but the Zipf-curve will not change much at all. For this reason, parameters a, b are better characteristic quantities for document frequences since they uniquely determine their distribution. a, b also have clear meaning: b is the exponent that governs the decay; a is the *expected* number of words with `df=1` according to the Zipf-curve. The fact that we can know the expected number of `df=1` words without actually counting `df=1` words indicates the value of the analysis of document frequence distribution.

Since document frequency is very often used as the global weighting for document representation in the vector space (as in tf.idf weighting), knowing their distribution will help to understand the effects of weighting and to further improve the weighting.

8.2 Distribution of Distinct Words

Next we investigate the number of distinct words (terms) in each document. This is the number of nonzero elements in each column in the term-document matrix X. In Figure 4, we plot the distribution of this quantity, i.e., the number of documents for a given number of distinct words. For the Cranfield collection, the minimum number of distinct words for a document is 11 (document # 506), and the maximum is 155 (document # 797). The peak point (40, 39) in the histogram indicates there are 39 documents in the collection, each of which has 40 distinct words.

This leads to a distribution very different from the Zipf distribution for document frequency above. The distribution appears to follow a log-normal distribution, which has the following probability density function

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma x} \exp\left(-\frac{(\log x - \mu)^2}{2\sigma^2}\right) \tag{17}$$

with mean and variance:

$$E(x) = e^{\mu+\sigma^2/2}, \; var(x) = e^{2\mu+\sigma^2}(e^{\sigma^2} - 1). \tag{18}$$

Calculating the mean E and variance var directly from the data, we can solve Eq.(18) to

obtain the parameters μ, σ. The probability density function can be drawn (the smooth curves in Figure 3). This simple procedure provides a very good fit to the data.

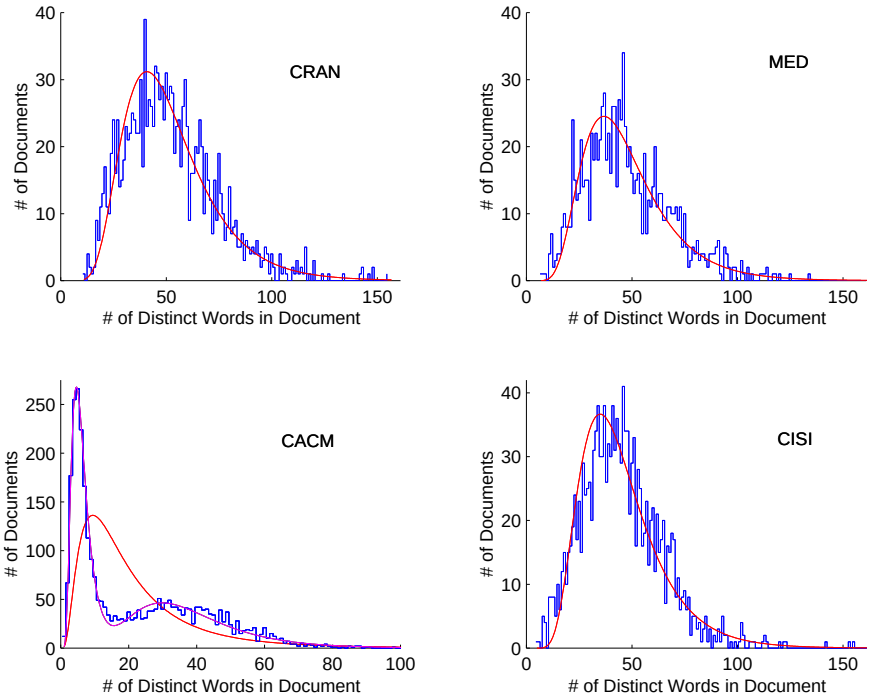

Figure 4: Distribution of distinct words for four document collections. Smooth curves are from the log-normal distribution (see text).

For three document collections, Cranfield, Medline, and CISI, the number of distinct words follow log-normal distributions. Normally, we expect this quantity to follow a normal (Gaussian) distribution. However, due to the fact that the quantity is a non-negative count variable, we expect the logarithm of the variable to follow a Gaussian distribution with mean μ and variance σ^2. This leads to the log-normal distribution for the original variable.

At first look, the histogram for CACM seems to deviate substantially from the log-normal distribution, as shown by the smooth single peak curve determined by the mean and variance of the entire data points in CACM. However, a careful examination of the two-peak pattern indicates that each of them can be fitted by a log-normal distribution. In fact, the second peak part is quite similar to all three other collections, with the peak number of documents occurs at around 40 distinct words, and with similar magnitude of vertical scale.

A simple fit of the CACM histogram with two log-normal distributions $p(x) = w_1 p_1(x) +$

$w_2 p_2(x)$ is also shown in Figure 3, with

$$\mu_1 = 1.75, \sigma_1 = 0.472$$

and

$$\mu_2 = 3.57, \sigma_2 = 0.415.$$

The fit is quite good. From the weights w_1, w_2, we can calculate the number of documents in each log-normal distribution. The calculated values are 1633 docs for the left peak part, and 1571 docs for the right peak part.

The CACM collection consists of titles and abstract of articles in *Communications of ACM*. However, out of the 3204 CACM documents, 1617 documents contain titles only, and therefore have much less number of distinct words per document (around 5). The remaining 1587 documents contain both a title and an abstract, therefore have number of distinct words around 30-40. Our statistical analysis automatically picks up the substantial difference and gives very close estimates of documents in each category: 1633 vs 1617 for the title-only documents and 1571 vs 1587 for the title+abstract documents. This indicates the usefulness of statistical analysis of document collections.

9 Related Work

With the contexts and notations provided above, we give pertinent descriptions of related work.

Traditional IR probabilistic models, such as the binary independence retrieval model [Rijsbergen, 1979, Fuhr, 1992] focus on relevance to queries. There, relevance to a specific query is pre-determined or iteratively determined in the relevance feedback, on individual query basis. Our new approach focuses on the term-document matrix using a probabilistic generative model. This occurrence probability could also be used in the language modeling approach for IR [Ponte and Croft, 1999].

Similarity matrices XX^T and X^TX are key considerations of our model. X^TX is used as the primary goal in the multi-dimensional scaling interpretation[6] of LSI where it is shown that LSI is the best approximation to X^TX in the reduced k-dimensional subspace. There,

the document-document similarity is also generalized to include arbitrary weighting, which improved the retrieval precision.

If the first k singular values of SVD are well separated from the rest, the k-dim subspace is proved to be stable against smaller perturbations in [Papadimitriou et al, 1998, Azar et al, 2000]. A probabilistic corpus model built upon k topics is then introduced and is shown to be essentially the LSI subspace[Papadimitriou et al, 1998].

A subspace model using the low-rank-plus-shift structured is introduced in [Zha et al, 1998] and lead to a relation to determine optimal subspace dimension $k_{\rm int}$ from the singular values. The relation was originally developed for array signal processing using minimum description length principle.

Using a spherical K-means method for clustering documents [Dhillon and Modha,2001] leads to the concept vectors (centroids of each clusters), which are compared to LSI vectors. The subspace spanned by the concept vectors are close to the LSI subspace. This method is further developed using efficient clustering algorithms into concept indexing [Karypis and Han, 2000].

Introducing a latent class variable in the joint probability for P(document,word), the resulting probability of the aspect model [Hofmann, 1999] can be written as $U\Sigma V^T$ of the SVD. This *direct* probabilistic formalism is further developed to handle queries and its effectiveness is shown. The latent class and the concept vector/index have many common features.

We briefly mention two further improvements of LSI. Using a coordinate rotation $R = V_k^T$ (see footnote in section 3), the LSI k-dim subspace can be see as spanned by $U_k V_k^T$ which is equivalent to set $\sigma_1 = \cdots = \sigma_k = 1$. This "dimension equalization" of LSI is further developed into trans-lingual method that appears to work well for large corpas [Jiang and Littman, 2000]. Another development is iterative scaling of LSI, which appears to improve the retrieval precision [Ando, 2000].

An important advantage of LSI is the reduced storage requirements on the database, by storing only a small number of singular vectors, rather than the original term-document matrix. However, since term-document matrices in information retrieval are usually very sparse ($< 1\%$ nonzero), the storage savings is not significant. Several recently developed methods further significantly reduce the storage of singular vectors by either using a dis-

crete approximation [Kolda and O'Leary, 1998] or thresholding on values of the LSI/SVD vectors[Zhang, 1999].

10 Discussions

We have outlined the dual model and worked out a few results. Here we point out several more features of the model.

10.1 Invariance Properties

The model has several invariance properties. First, the model is invariant with respect to (w.r.t.) the order that words or documents are indexed, since they depend on the dot-product which is invariant w.r.t. the order. The singular vectors and values are also invariant, since they depends on XX^T and X^TX , both of which are invariant w.r.t. the order.

Second, the similarity relations between documents and between words are preserved in the k-dim LSI subspace. In the LSI subspace, documents are represented as their projections, i.e., columns of $U_k^T X = \Sigma_k V_k^T$; words are represented as the rows of $XV_k = U_k\Sigma_k$. The document-document similarity matrix in the LSI subspace is

$$(\Sigma_k V_k^T)^T(\Sigma_k V_k^T) = (U_k\Sigma_k V_k^T)^T(U_k\Sigma_k V_k^T) \simeq X^TX, \tag{19}$$

up to the minor difference due to the truncation in SVD. Similarly, the term-term similarity matrix in LSI subspace is

$$(U_k\Sigma_k)(U_k\Sigma_k)^T = (U_k\Sigma_k V_k^T)(U_k\Sigma_k V_k^T)^T \simeq XX^T, \tag{20}$$

up to the minor difference due to the truncation in SVD.

Note that the self similarities, i.e., the diagonal elements in the similarity matrices, are the length of the vectors (L_2 norm). Thus, if document vectors are normalized in the original space, they remain approximately normalized in the LSI subspace. For this reason, we believe that documents should be normalized before LSI is applied, to provide a consistent view.

Third, the model is invariant with respect to a scale parameter s, an average similarity,

which could be incorporated in Eq.(4) as,

$$\Pr(\mathbf{x}_i|\mathbf{c}_1 \cdots \mathbf{c}_k) \propto e^{[(\mathbf{x}_i\cdot\mathbf{c}_1)^2+\cdots+(\mathbf{x}_i\cdot\mathbf{c}_k)^2]/s^2}, \tag{21}$$

similar to the standard deviation in Gaussian distributions. We can repeat the analysis in Section 5, and obtain the same LSI dimensions and same likelihood curves except that the vertical scale is enlarged or shrinked depending on $s > 1$ or $s < 1$.

10.2 Normalization Factors

In this paper, we started with documents (columns of X) that are normalized: $||\mathbf{x}_i|| = 1$. This implies that the cosine similarity is equivalent to the dot-product similarity between documents:

$$\mathrm{sim}_{\cos}(\mathbf{x}_1, \mathbf{x}_2) = \mathbf{x}_1 \cdot \mathbf{x}_2/||\mathbf{x}_1|| \cdot ||\mathbf{x}_2|| = \mathrm{sim}_{\mathrm{dot}}(\mathbf{x}_1, \mathbf{x}_2).$$

However, the normalization of columns does not imply that each term (rows of X) are normalized to $\sqrt{n/d}$ (see Eq.15), although they do so on average.

This implies that for words, the dot-product similarity is not equivalent to cosine similarity. This is not a serious problem in itself, since dot-product similarity is a well-defined similarity measure. But, the symmetry between documents and words is not entirely preserved here.

Can the rows and columns be normalized simultaneously? The answer is affirmative. We can achieve this by the following simple iterative procedure, i.e., alternatively normalizing rows and columns. This procedure, however, is not uniquely defined: Given an un-normalized X, we can either (a) first normalize the columns and then normalize the rows, or (b) first normalize the rows and then normalize the columns. The results in these two cases may differ within the first few iterations.

We can prove the following simple theorem: the final result for simultaneous normalizations of rows and columns is unique under general conditions (see Appendix for the proof). Therefore, even though the results after a few iterations may be different depending on whether the columns are first normalized or not, when the process converges, the final results is independent of the initial steps. We did experiments on few matrices and verify the correctness of the theorem. Therefore, the columns and rows can both be normalized to a

constant simultaneously. Afterwards, for terms $\mathbf{t}^1, \mathbf{t}^2$ we have

$$\text{sim}_{\cos}(\mathbf{t}^1, \mathbf{t}^2) = \mathbf{t}^1 \cdot \mathbf{t}^2 / ||\mathbf{t}^1|| \cdot ||\mathbf{t}^2|| = (n/d) \cdot \mathbf{t}^1 \cdot \mathbf{t}^2 = (n/d) \cdot \text{sim}_{\text{dot}}(\mathbf{t}^1, \mathbf{t}^2),$$

showing that the dot-product similarity is the same as the cosine similarity (the proportional constant (n/d) will not change ranking and is thus irrelevant).

10.3 Separation of Term and Document Representations

In the LSI subspace, documents and words are are represented by their projections. The dual relationship between them is no longer directly represented as rows and columns of the *same* matrix. Instead, they are related through a *filtered* procedure, $\mathbf{u}_j = (1/\sigma_j) X \mathbf{v}_j$. This is similar to a learning process: from several contexts, the meaning of a word is better described by a number of filtered contexts, instead of the original raw contexts.

11 Summary

In this paper, we introduced a dual probabilistic generative model based on similarity measures. Similarity matrices then arise naturally during the maximum likelihood estimation process, and LSI is the optimal solution of the model.

Semantic associations characterized by the LSI dimensions are measured by their statistical significance, the likelihood. Calculations on four standard document collections exhibit a maximum in likelihood curves, indicating the existence of an intrinsic semantic subspace. The importance of LSI dimensions follows a Zipf-like distribution, indicating LSI dimensions represent distinct identities or latent concepts.

The term-document matrix is the main focus of this study. The number of nonzero elements in each row of the matrix is the document frequency, which follows Zipf-distribution. This is the direct reason that the statistical significance of LSI dimensions follow the Zipf law. The number of nonzero elements in each column of the matrix is the number of distinct words, which follows log-normal distribution and gives useful insights to the structure of the document collection.

Besides automatic information retrieval, text classification, and word sense disambiguation, our model can apply to many other areas, such as image recognition and reconstruction, as long as the relevant structures are essentially characterized or defined by the dot-product similarity.

Overall, the model provides a statistical framework upon which LSI and similar dimension reduction methods can be analyzed and further improved. For example, in the present form, queries do not enter the picture; a further development involving queries explicitly could yield insights into the information retrieval process. Very recently, an iterative scaling method improved upon LSI is proposed [Ando, 2000] that utilized our model in analyizing the processes involved there.

Beyond information retrieval and computational linguistics, LSI is used as the basis for a new theory of knowledge acquisition and representation[Landauer and Dumais, 1997] in cognitive science. Our results that LSI is an optimal procedure and that the intrinsic semantic subspace is much smaller than the initial semantic space lend deeper understanding and support to that theory.

Acknowledgements. The author wishes to thank Dr. Hongyuan Zha for providing term-document matrices used in this study and for initiating the IR research, Drs. Zhenyue Zhang, Osni Marques, Parry Husbands and Horst Simon for valuable discussions, Drs. Micheal Berry and Inderjit Dhillon for seminars given at NERSC/LBL that help motivated this work, and Dr. Susan Dumais for communications. This work is supported by Office of Science, Office of Laboratory Policy and Infrastructure, of the U.S. Department of Energy under contract DE-AC03-76SF00098.

Appendix.

Here we prove that the final result for simultaneous normalizations of rows and columns of a general matrix is unique. For this, we introduce weight factors $(a_1, a_2, \cdots, a_n)$ and $(b^1, b^2, \cdots, b^d)$, such that every column is normalized

$$\sum_{\alpha=1}^{d} (a_i b^\alpha x_i^\alpha)^2 = 1, \quad i = 1, \cdots, n \tag{22}$$

and every row is normalized at same time:

$$\sum_{i=1}^{n} (a_i b^\alpha x_i^\alpha)^2 = n/d, \quad \alpha = 1, \cdots, d \tag{23}$$

where x_i^α are elements of the word-to-document matrix X. These $n+d$ quadratic equations for variables $\{a_i\}, \{b^\alpha\}$ can be solved using an iterative procedure, by alternatively normalizing rows and columns.

We prove that the solution is unique. Suppose we have two distinct sets of solutions, $\{a_i, b^\alpha\}$ and $\{\tilde{a}_i, \tilde{b}^\alpha\}$. Substitute them into Eqs.(22,23) and subtract one from another, we have

$$\sum_{\alpha=1}^{d} (x_i^\alpha)^2 y_i^\alpha = 0, \quad i = 1, \cdots, n \tag{24}$$

and

$$\sum_{i=1}^{n} (x_i^\alpha)^2 y_i^\alpha = 0, \quad \alpha = 1, \cdots, d. \tag{25}$$

Now we view $\{a_i\}, \{b^\alpha\}$ as given from an known solution, and solve for $n+d$ variables $\{\tilde{a}_i\}, \{\tilde{b}^\alpha\}$ through the combination $y_i^\alpha = (a_i b^\alpha)^2 - (\tilde{a}_i \tilde{b}^\alpha)^2$ from the $n+d$ homogeneous equations. Under general conditions, the determinants of the coefficient matrices of these equations [made out of different combinations of $(x_i^\alpha)^2$] will not be zero. Thus $y_i^\alpha = 0$ for all i, α. This implies that $(a_i)^2 = (\tilde{a}_i)^2$ for all i and $(b^\alpha)^2 = (\tilde{b}^\alpha)^2$ for all α. Therefore the two solutions must be identical: the final solution is unique.

References

[Ando, 2000] R. K. Ando. Latent Semantic Space: Iterative Scaling Improves Precision of Inter-document Similarity Measurement. Proc. SIGIR-2000, 2000, ACM Press, pp.216-223.

[Azar et al, 2000] Y. Azar, A. Fiat, A. Karlin, F. McSherry and J. Saia. Spectral analysis for data mining. In Proc. 33rd Annual ACM Symposium on Theory of Computing (STOC), 2000.

[Baker and McCallum, 1998] L.D. Baker, A.K. McCallum. Distributional Clustering of Words for Text Classification. Proc. SIGIR-98, New York, 1998. ACM Press.

[Bartell et al,1995] B.T. Bartell, G.W. Cottrell, and R.K. Belew. Representing Documents Using an Explicit Model of Their Similarities. J.Amer.Soc.Info.Sci, **46**, 251-271, 1995.

[Berry et al 1995] M.W. Berry, S.T. Dumais, G.W. O'Brien. Using linear algebra for intelligent information retrieval. SIAM Review, **37**, 573 (1995).

[Caron, 2000] J. Caron. Experiments with LSA Scoring: Optimal Rank and Basis. Proc. of SIAM Computational Information Retrieval Workshop, October 2000. Ed. M. Berry.

[Deerwester et al, 1990] S. Deerwester, S.T. Dumais, T.K. Landauer, G.W. Furnas, R.A. Harshman. Indexing by latent semantic analysis. J.Amer.Soc.Info.Sci, **41**, pp:391-407 (1990).

[Dhillon and Modha,2001] I. Dhillon and D. Modha. Concept Decomposition for Large Sparse Text Data Using Clustering. *Machine Learning.* **42**, pp.143-175. (2001).

[Ding, 1999] C.H.Q. Ding. A Similarity-based Probability Model for Latent Semantic Indexing. Proc. of SIGIR'99 (ACM Press, 1999), pp.59-65.

[Dumais, 1995] S.T. Dumais. Using LSI for information filtering: TREC-3 experiments. Third Text REtrieval Conference (TREC3), D. Harman, Ed., National Institute of Standards and Technology Special Publication, 1995.

[Fuhr, 1992] N. Fuhr. Probabilistic Models in Information Retrieval. Computer Journal, **35**, 243 (1992).

[Furnas et al, 1987] G.W. Furnas, T.K. Landauer, L. Gomez, S.T. Dumais. The Vocabulary Problem in Human-system communications. Commun. ACM, **30**, pp:964-971. (1987).

[Glassman, 1994] S. Glassman. A Caching Relay for the World Wide Web. Comput. Networks ISDN System, **27**, 165 (1994).

[Golub and Reinsch, 1971] G. Golub and C. Reinsch. Handbook for Automatic Computation II, Linear Algebra. Springer-Verlag, New York, 1971.

[Golub and Van Loan, 1989] G. Golub and C.V. Loan. Matrix Computation. Johns-Hopkins, Baltimore, 2nd ed. 1989.

[Hofmann, 1999] T. Hofmann. Probabilistic Latent Semantic Indexing. Proc. of SIGIR'99 (ACM Press, 1999), pp.50-57.

[Husbands et al, 2000] P. Husbands, H. Simon and C.Ding. On the Use of Singular Value Decomposition for Text Retrieval. Proc. of SIAM Comp. Info. Retrieval Workshop, October 2000. Ed. M. Berry.

[Jiang and Littman, 2000] F. Jiang and M.L. Littman, Approximate dimension equalization in vector-based information retrieval. Proc. 17th Int'l Conf. Machine Learning, 2000.

[Karypis and Han, 2000] G. Karypis and E.H. Han. Concept indexing: A fast dimensionality reduction algorithm with applications to document retrieval and categorization. Proc. 9th Int'l Conf. on Info. Knowledge Management (CIKM 2000).

[Katz, 1996] S.M. Katz. Distribution of content words and phrases in text and language modeling. Natural Language Engineering, Vol.2, pp.15-60 (1996).

[Kolda and O'Leary, 1998] T.G. Kolda and D.P. O'Leary. A Semi-Discrete Matrix Decomposition for Latent Semantic Indexing in Information Retrieva. ACM Trans. Information Systems, 16 (1998), pp.322-346.

[Landauer and Dumais, 1997] T.K. Landauer and S.T. Dumais. A Solution to Plato's Problem: the Latent Semantic Analysis Theory of Acquisition, Induction and Representation of Knowledge. Psychological Review, **104**, 211-240 (1997).

[Manning and H. Schütze, 1999] C.D. Manning and H. Schütze. Foundations of Statistical Natural Language Processing. MIT Press. Cambridge, MA. 1999.

[Papadimitriou et al, 1998] C.H. Papadimitriou, P. Raghavan, H. Tamaki, S. Vempala. Latent Semantic Indexing: A Probabilistic Analysis. Proc. of Symposium on Principles of Database Systems (PODS), Seattle, Washington, June 1998. ACM Press.

[Ponte and Croft, 1999] J. Ponte and W.B. Croft. A Language Modeling Approach to Information Retrieval. Proc. of SIGIR'98 (ACM Press, 1999), pp.275-281.

[Rijsbergen, 1979] C.J. van Rijsbergen. Informational Retrieval. 2nd Ed. (Butterworths. 1979).

[Salton and McGill, 1983] G. Salton and M.J. McGill. Introduction to Modern Information Retrieval. (McGraw-Hill, 1983).

[Salton and Buckley, 1988] G. Salton and C. Buckley. Term-weighting approaches in automatic text retrieval. Information Processing and Management, **24**, pp.513-524 (1988).

[Schütze, 1998] H. Schütze. Automatic Word Sense Discrimination. Computational Linguistics, 24, pp:97-124, 1998.

[Story, 1996] R.E. Story. An explanation of the effectiveness of latent semantic indexing by means of a Bayesian regression model. Information Processing & Management, 32, pp.329-344. 1996.

[Yang, 1995] Y. Yang. Noise Reduction in a Statistical Approach to Text Categorization. Proc. of SIGIR'95 (ACM Press, 1995), pp.256-263.

[Zha et al, 1998] H. Zha, O. Marques and H. Simon, A Subspace-Based Model for Information Retrieval with Applications in Latent Semantic Indexing. Proc. of Irregular '98, Lecture Notes in Computer Science, **1457**, (Springer-Verlag, 1998), pp.29-42.

[Zhang, 1999] Z. Zhang, H. Zha, and H. Simon. Low-Rank Approximations with Sparse Factors I: Basic Algorithms and Error Analysis. July 1999. Submitted to SIAM Journal of Matrix Analysis.

[Zipf, 1949] G.K. Zipf. Human Behavior and the Principle of Least Effort. Addison-Wesley, 1949.